

Open Repositories 2009

Integrating Fedora into DiVA

Uwe Klosa, Electronic Publishing Centre, Uppsala University

Proposal for a presentation for OR09 - Fedora User Group

Abstract

This presentation will provide an insight into the process of integrating Fedora into DiVA. It will describe how Fedora fits into the system structure of DiVA, and how its services are employed to join together the various modules.

From the start, DiVA had made extensive use of XML for storing metadata in files — in a repository specifically developed for DiVA — together with full-text files, images, videos and so on. Additionally, in conjunction with XSL, DiVA had utilised XML for the representation of its content.

When we decided to rebuild DiVA from scratch, one question of concern was which repository software to use. The decision to use Fedora as the repository software for the new DiVA was taken after the first beta version of Fedora 3.0 had been released.

In Fedora we created a basic content model for digital publication objects, which is mainly employed as a straightforward aid to provide a number of services, such as the dissemination of various metadata formats in XML. In publication objects, the metadata are stored as internal data streams. All files belonging to a publication are stored as managed content data streams. In addition, version control and checksumming are used to document changes and to recover an older version if necessary.

A storage client in DiVA accesses Fedora via its SOAP interfaces. The storage client provides the option to create and update objects in the repository. A different client with read-only access is used to deliver files to the portal and local web pages.

The content in the repository is indexed by several instances of Apache SOLR which create distinct indexes for different purposes. One index for the administrative interface, another one for the web pages, and a third one for DiVA's OAI data provider and search services. The SOLR instances have their own clients who listen to messages from Fedora's messaging service.

Currently DiVA stores over 200.000 publication objects, around 33.000 of which contain full-text files and other attachments. The indexes are roughly 20G in total and the repository is about 115G. The whole DiVA system runs on Solaris with 3 Zones and Ubuntu Xen with 13 virtual machines.

